Using Adaptive Comparative Judgment in Writing Assessment: An Investigation of Reliability Among Interdisciplinary Evaluators

By Sweta Baniya, Nathan Mentzer, Scott R. Bartholomew, Amelia Chesley, Cameron Moon and Derek Sherman

ABSTRACT

Adaptive Comparative Judgment (ACJ) is an assessment method that facilitates holistic, flexible judgments of student work in place of more quantitative or rubric-based methods. This method "requires little training, and has proved very popular with assessors and teachers in several subjects, and in several countries" (Pollitt 2012, p. 281). This research explores ACJ as a holistic, flexible, interdisciplinary assessment and research tool in the context of an integrated program that combines Design, English Composition, and Communications courses. All technology students at Polytechnic Institute at Purdue University are required to take each of these three core courses. Considering the interdisciplinary nature of the program's curriculum, this research first explored whether three judges from differing backgrounds could reach an acceptable level of reliability in assessment using only ACJ, without the prerequisites of similar disciplinary backgrounds or significant assessment experience, and without extensive negotiation or other calibration efforts. After establishing acceptable reliability among interdisciplinary judges, analysis was also conducted to investigate differences in student learning between integrated (i.e., interdisciplinary) and non-integrated learning environments. These results suggest evaluators from various backgrounds can establish acceptable levels of reliability using ACJ as an alternative assessment tool to more traditional measures of student learning. This research also suggests technology students in the integrated/interdisciplinary environment may have demonstrated higher learning gains than their peers and that further research should control for student differences to add confidence to these findings.

Keywords: Adaptive Comparative Judgement (ACJ), assessment, interdisciplinary learning/environment, interdisciplinary evaluators, integrated STEM education, engineering and technology, design thinking, composition, and communication.

INTRODUCTION

Strong writing, speaking, and critical thinking skills are valuable 21st century competencies in high demand for graduates of technology programs. With present day educational trends becoming more interdisciplinary in nature, "judgments of proficiency must also be made on the basis of performances in multiple and varied writing situations (for example, a variety of topics, audiences, purposes, genres)" ("CCCCs Position Paper on Writing Assessment, 2014). Several researchers have discussed the value of interdisciplinary teaching approaches within STEM disciplines and beyond (Bannerot, Kastor, & Ruchhoeft, 2010; Wang, Moore, Roehrig, & Park, 2011). Kellam, Walther, Constantio, Dodd, & Cramond (2013), for example, described a four-year engineering integration program in which students are encouraged toward "develop[ing] a deep understanding of the larger socio-technical systems in which engineering is situated" (p. 8). However, assessment of the complex, situated writing and communication skills such integration programs aim to promote can be particularly challenging for instructors and program administrators. Given the increasingly interdisciplinary technology degree programs and classrooms, how can instructors and program administrators meaningfully and reliably evaluate students' situated, interdisciplinary learning?

The context for this study involved an interdisciplinary pedagogical effort in which program leaders from the courses of English Composition, Communications, and Design Thinking in Technology linked several sections to teach cohorts of students in an integrated environment. Technology students at Purdue University are required by their college to take the introductory Design Thinking course and by the university to take English Composition and Communications courses. This course integration program was an engaged effort to teach design and technology alongside introductory composition and communication. The overarching program and its assessment efforts have aimed to understand if and how technology students learn to communicate more

effectivity through integrated pedagogy. As the evaluation of the effectiveness of the integrated program began, assessors recognized that in an interdisciplinary learning environment, assessment of student learning requires a holistic approach that can flexibly compare multiple genres of writing produced in multiple educational contexts having multiple purposes, and that is both valid and reliable (Moss 1994; White 1984). An interdisciplinary assessment should be a varied and flexible method of assessment that can incorporate a variety of student compositions produced based on program requirements and relevant composition pedagogy (Gallagher 2014).
Yet the ongoing growth of interdisciplinary courses comes with deep uncertainty about how to structure and evaluate interdisciplinary learning experiences and measure student success (Mansilla, Duraisingh, Wolfe, & Haynes, 2009). In order to understand the impact of technology students' interdisciplinary experiences on their writing abilities, this study set up and implemented a holistic interdisciplinary assessment of students' work within an integrated classroom. Discussion of the integration program and the non-integrated sections used for comparison in the research will be discussed in detail in methods section. To investigate a new method of holistic evaluation and find out what it revealed about students' learning, evaluators from different fields were assembled to evaluate submissions with the question: "Can these three evaluators develop acceptable levels of agreement on the relative quality of student work?" The major concern for conducting this research was that the diversity of evaluators' perspectives, so essential for promoting interdisciplinary thinking and learning, might also negatively affect their levels of agreement and thus jeopardize the reliability and validity of the results.

This article uses the term interdisciplinary to refer to any collaboration or environment whereby individuals or principles from multiple academic disciplines join for a common purpose or project. Thus "interdisciplinary evaluators" means a group of evaluators with differing disciplinary expertise and "interdisciplinary learning environment" means a space in which multiple disciplinary principles are discussed and explored.

This article begins with a review of major writing assessment scholarship that highlights the need for interdisciplinary and holistic writing assessment and explores the potential of Adaptive Comparative Judgment (ACJ) as a holistic assessment tool and process. Next, the article reviews the process by which Adaptive Comparative Judgment (ACJ) has been implemented for this research via a digital interface called Compare Assess and describes this method in detail, along with its use in context of an interdisciplinary course integration program. Three evaluators were invited to assess a sample of student writing to investigate whether they could establish acceptable levels of inter-rater reliability without extensive calibration efforts and to determine whether there were significant differences between artifacts collected from the integrated course and those collected from the non-integrated sections. Following a description of the study's context is a presentation and discussion of the research concerning Adaptive Comparative Judgement as a reliable tool for writing assessment in the context of an integrated, interdisciplinary classroom.

Contextualizing Writing Assessment Literature

The sections of the following literature review focus on writing assessment broadly before turning toward scholarship on the application of Adaptive Comparative Judgment (ACJ) more specifically. The reason for focusing here on writing assessment literature is to create an understanding of how writing assessment strategies could be used to assess writings of technology students. This literature review also allows us to think about the gaps that exist in writing assessment theory and practice, and about how established strategies may fail to address the assessment needs of multiple educational contexts and disciplines.

Holisticism in Writing Assessment

Holisticism is an approach to assessment which measures the student's ability of writing, considering it as a "whole" and not as multiple "parts." Holisticism in writing assessment, according to White (1984), was first introduced by the Educational Testing Services (ETS). Lynne (2004) added that test development specialists at the ETS were the ones who devised holistic scoring in response to their smaller clients who wanted to see actual pieces of writing and who were less concerned with efficiency. With changing demands on writing programs, the growth of interdisciplinary pedagogies, and new approaches in writing instruction, writing assessment requires more dynamic and holistic measures to assess student writing. White (1984) argued that holisticism is a reliable, valid, and meaningful way of writing assessment; this approach encourages taking a general impression of students' writing, which means writing is evaluated as a whole, without any strict guides or controls. Holistic writing assessment also focusses on the idea of sensitivity, reader calibration, and creation of a positive social environment among the readers.

As English assessment philosophy matured, Yancey (2012) described three different waves of assessment and argued that the current trends of writing assessment were in flux. The first wave was mostly focused on "machinelike efficiency" and developing low-cost and fair measures (Yancey, 2012, p. 168), often represented in multiple choice assessments that led assessors to higher reliability. The second wave "was prompted, at least in part, by the explosion of interest in writing process and in new pedagogies enacting the field's new understanding of process as well as evolution of the holistic scoring" (Yancey, 2012, pp. 168-169). This second wave was characterized by a holistic way of assessment and primary trait scoring.

In contrast, the third wave of writing assessment focuses on writing portfolios and has been "characterized by attention to multiple texts, the ways [individuals] read those texts, and the role of students in helping [to] understand their texts and the processes they used to produce them" (Yancey, 2012, p.169). Portfolio assessment allows assessors to adapt new, evolving assessment theories and contextualize assessment based on programmatic needs. Moreover, Yancey (2012) argued that current assessment is complicated and dynamic because of the changes in teaching and assessing writing. This current wave recognizes multiple contexts in writing assessments with writing assessors taking all contexts into account, because "learning outcomes are linked with assessment with an emphasis on the programmatic assessment and also with a criticism of previous assessment practices" (Yancey, 2012, pp.172-173).

Need for New Writing Assessment

While Yancey stated that assessment is in flux, Huot (2002) argued for "reimagining writing assessment as a positive force in the teaching of writing" (p. 2). Huot proposed that researchers, teachers, and administrators need to create an understanding of writing assessment that is socially progressive and serves the purpose of teaching and learning. Holistic assessment, Huot argued, "provides models for assignment construction, fair grading practices, and the articulation of clear course goals" (p. 32). Holistic writing assessment is "able to achieve acceptably high reliability by adding a series of constraints to the economically efficient practice of general impression scoring" (White, 1984, p. 403).

While holistic writing assessment was developed about four decades ago, current writing assessment, as Huot (2002) stated, should be rearticulated, reimagined, and should meet the requirements of interdisciplinary classrooms. In the case of this research, interdisciplinary assessment has been implemented in an integrated, interdisciplinary classroom that combines technology-based study with composition and communication. Writing assessments not only help establish a clear understanding of whether a writing program is effective, but they also help in creating interdisciplinary exchange and stronger communication among various concerned stakeholders. Similarly, these approaches assist in identifying what aspects of a program should be enhanced and what learning goals should be increased such that the writing will create a positive social action.

While Huot (2002) pushed for a re-articulation of writing assessment and deeper thinking about the larger impacts of writing assessment, Gallagher (2014) presented a new perspective regarding reliability. One of the important aspects that White (1984), Yancey (2012), and Huot (2002) all discussed is how new approaches to writing assessment can be made to better fit the needs of programs and students' learning. In this regard, Gallagher (2014) explored three reliability theories in writing assessment: positivist, hermeneutic, and rhetorical, and proposed that writing assessment should practice rhetorical reliability. In comparison to other modes of reliability, Gallagher stated his belief that rhetorical reliability can reinvigorate the assessment work and allow the assessors to frame reliability in ways that articulate and advance rhetorical understandings of writing and writing assessment (p. 74). A focus on rhetorical reliability allows assessors to think about an artifact rhetorically and helps them understand whether it meets programmatic goals related to writing. However, often—especially for English professionals, as Huot (2002) argued—writing assessors "have been made to feel inadequate and naïve by considerations of technical concepts like validity and reliability" (p. 81). Thinking about reliability rhetorically as a result or common ground is necessary in cases and scenarios where all evaluators evaluate with a focus on rhetorical awareness of their process as well as of the artifact.

Comparative Judgment in Writing Assessment

Adaptive Comparative Judgment (ACJ) could be proposed as a holistic approach to assessment, which builds on the evolution of writing assessment and is a method that can lead to the kind of rhetorical reliability proposed by Gallagher (2014). However, before considering ACJ, it is necessary to understand "Comparative Judgment." Pollitt (2012) recounted that "the method of comparative judgement (CJ) was originally proposed by Louis L. Thurstone for work in psychophysics" (p. 282). Comparative Judgment includes making comparison between two artifacts rather than subjectively assessing one artifact at a time, and has been used in various disciplines, including writing, design, technology, engineering, pedagogy, and foreign language studies (Bartholomew & Yoshikawa, 2018). Scholars in Europe, especially in education, have been utilizing CJ to focus on efficient, low-labor, low-cost, and reliable assessment of writing. There have been some successful assessment results in Europe like that of Pollitt (2012) and in the United States. by scholars like Bartholomew, Strimel & Yoshikawa (2018) who have successfully shown ACJ to be useful in multiple educational contexts and disciplines.

In their article, Steedle and Ferrara (2016) suggested that comparative judgment could be used for assessing writing holistically. Likewise, in another study, Van Daal, Lesterhuis, Coertjens, Donche, & De Maeyer, (2019) suggested that "CJ enables judges to differentiate between essays on characteristics rooted in the quality of the essays allowing flexibility" (p. 11). Further, they concluded "the method is not only able to generate reliable scores, but also provided valid scores with regard to academic writing, even when a community of assessment practice is lacking" (p. 14). Both of these studies showed success in the use of CJ in writing assessment and propose CJ as a reliable, cost-efficient and valid method of writing assessment challenging the traditional notion of importance of creation of community in writing.

Adding "Adaptive" to Comparative Judgment

Based on the Comparative Judgement method, Pollitt (2012) proposed "Adaptive Comparative Judgment" as an alternative assessment approach. Much like CJ, "ACJ is a method of scoring students' work in which judges are asked only to make a choice between two student artifacts and decide which one is better" (Pollitt, 2012, p. 297). Unlike CJ, ACJ applies an algorithm to adaptively pair similarly judged items – thus potentially reducing the number of overall judgments required for a reliable rank order to be achieved. According to Pollitt (2012), the real power of CJ is "only realized when it is made adaptive" (p. 284). A web-based application was created by the Digital Assess company to facilitate the adaptive process of judgment using ACJ with a user-friendly interface. The application was made accessible via a web-portal called Compare Assess. The output of an ACJ session conducted using Compare Assess typically consists of several items including a rank order of included items and a parameter value for each item. The parameter values represent the relative quality of each item and thus differ from the rankings; the parameter values may be close or distant depending on the relative quality of items when compared to others (see Pollitt, 2012; 2015). Output also includes Rasch's misfit statistics, which can be used to identify judges or items that demonstrated significant difference or difficulty in agreement.

Pollitt (2012) also presented the results of a pilot study conducted to assess the writing of students in England aged 9-11 years using ACJ. In this pilot study, 54 evaluators were used (31 were professional test/examination markers and the rest were teachers who didn't have training on marking). The evaluators were trained on the use of the web-based program but were not provided training for the calibration in assessment (Pollitt, 2012). While the focus of their research was to test whether ACJ would be successful for assessment purposes, they did not focus on testing interdisciplinary evaluators for assessment. To fill this gap, the present study investigates whether a group of interdisciplinary evaluators can evaluate interdisciplinary work with high levels of reliability, or if their different value structures would result in disagreements and low reliability.

Motivation and Context for the Current Study

The integrated learning environment used as the context of this study involved a program developed for first-year technology students. With a goal of supporting an interdisciplinary English composition pedagogy, strengthening design thinking skills and enhancing communications skills, Purdue University introduced an "Integrated First-Year Experience" program, providing students with opportunities to connect and transfer knowledge from various disciplines in collaborative contexts that simulate contemporary, 21st-century workplaces.

The overall goals of the Integrated First-Year Experience program are to create an intersection where students foster design thinking and critical thinking, enhance rhetorical awareness, and develop oratorical skills. These intended learning outcomes were established based on the programmatic goals and outcomes of programs within three academic units: the Polytechnic Institute, the Department of English, and the Department of Communication. Furthermore, the integrated curriculum efforts were inspired by the National Academy of Engineering publication on STEM Integration in K12 Education (Honey, Pearson, & Schweingruber 2014).

Freshmen students enrolled in this integrated program were placed in a class of 40-45 students, where all of them take a Design Thinking in Technology course (a core requirement for students in the Polytechnic Institute). Learning outcomes for this course are:

1. Write a narrowly focused problem statement;
2. Apply ethnographic methods to understand technological problems;
3. Develop a search strategy, access technical databases, and evaluate results and source quality;
4. Create a technical report documenting results of the design process;
5. Manage design projects, develop project timelines, and negotiate individual responsibilities and accountability in the team environment;
6. Apply strategies of ideation to develop novel and innovative solutions; and
7. Rapidly prototype solutions for purposes of design, testing, and communication.

Twenty of these students were also enrolled in the same section of Introductory Composition (required of nearly all undergraduates at Purdue University) where they are expected to:

1. Demonstrate rhetorical awareness of diverse audiences, situations, and contexts;
2. Compose a variety of texts in a range of forms, equaling at least 7,500-11,500 words of polished writing (or 15,000-22,000 words, including drafts),
3. Think critically about writing and rhetoric through reading, analysis, and reflection;
4. Provide constructive feedback to others and incorporate feedback into their writing;
5. Perform research and evaluate sources to support claims; and
6. Engage multiple digital technologies to compose for different purposes.

Likewise, 20-25 of the students simultaneously took Fundamentals of Speech Communication (also required of nearly all undergraduates) where students are expected to:

1. Effectively perform the role of the public speaker by learning principles of communication theory and how to apply those principles to the management of speaking situations both individually and in group presentations;
2. Demonstrate knowledge and skill in the following areas: Audience analysis, Topic analysis, Organizational skills, Persuasive and informative strategies, Verbal and non-verbal delivery skills, and Group communication skills; and
3. When making a presentation: Select an appropriate topic, Prepare a full sentence outline with bibliography, Provide appropriate transitions and summaries, Develop effective introductions and conclusions, Use an appropriate organizational pattern, Use supporting material properly and effectively, Create effective presentational aids, Use presentational aids effectively and Display appropriate verbal and nonverbal behaviors.

In their Design Thinking in Technology class, students focused on designing solutions to a problem that is localized on the university's campus and had global implications aligned with the National Academy of Engineering's Grand Challenges ("National Academy of Engineering," n.d.). In Introductory Composition, students learned professional writing practices, developing analytical, critical, and research skills, and demonstrated these via multimodal compositions. Likewise, the Fundamentals of Speech Communication course aimed at enabling student's communication skills with focus on interpersonal communication and teamwork and presented informative and persuasive speeches.

Ideally, to make the integrated program successful, the instructors of all three courses worked collaboratively to identify and implement common goals and outcomes of the integrated classrooms. The instructors met regularly to foster partnership among themselves as well as pedagogical connections among their individual classroom teaching. Additionally, the program required collaborative, same-classroom co-teaching at multiple strategic points in the semester. In the second half of the semester, students worked on a collaborative integrated assignment that incorporated central elements from *Design Thinking, English Composition, and Speech Communication courses.* In the common final project, students implemented the knowledge they developed throughout the semester regarding design thinking, critical thinking, and clear communication by giving a team presentation at the end of the semester.

Prior Assessment Efforts

Prior to this research study, as part of a programmatic evaluation effort, the differences in student learning between the integrated sections and non-integrated sections of English Composition were explored. The research method employed was rubric-based, using a traditional, somewhat holistic, 6-point rubric for assessment of research writing. This rubric had been developed by the Department of English at Purdue University and has been used as an instrument for evaluation of the student papers throughout the first-year composition as well as for the purpose of departmental assessment and evaluation. Considering the use of rubrics as valid and reliable, two experienced composition instructors from the English department served as evaluators. After evaluating nearly 100 Composition submissions, some reflection on the process resulted in concerns that the interdisciplinary writing strengths that the researchers intended to measure were not being validly assessed by the general-purpose English rubric. The results from this previous effort suggest that the rubric-based assessment was a reductionist approach and was not able to adequately measure the rhetorical awareness and writing strengths of the interdisciplinary, multimodal, and multi-genre compositions.

With this result, a decision was made to review the same data with the more holistic ACJ assessment approach. Since ACJ is a comparative method, the control group from which half of the data for comparison was gathered included students in non-integrated English Composition sections, where students from any major or class standing were enrolled without planned concurrent enrollment in any of the other relevant courses. In addition to applying ACJ as the tool for holistic assessment, judges were invited from different disciplines to evaluate the artifacts. Research has not substantiated whether a diverse set of evaluators investigating writing across the disciplines can successfully establish acceptable levels of reliability, even though scholars like Pollitt (2012) have articulated the success of ACJ as an assessment method. With this study, researchers hope to recognize and demonstrate the value of integrated disciplines coming together in the assessment of writing.

Using evaluators from multiple disciplines can facilitate a more holistic assessment approach, though they may also introduce disagreement and complexities that disrupt the reliability of the measures. In previous research, evaluators shared disciplinary backgrounds and therefore calibration or reliability between them appeared to be a minimal issue (Bartholomew, Strimel, & Yoshikawa, 2018). However, this research introduces diversity in the assessment team, which provides an opportunity to test whether such a diverse team would be capable of agreeing on what they considered strong student writing. Furthermore, another purpose of this research is to show whether the diverse team would be reliable without costly and extensive pre-assessment training and calibration efforts. Based on these purposes, this research attempts to answer two research questions related to holistic and interdisciplinary assessment:

1) Can an ACJ approach to holistic English composition assessment be implemented by interdisciplinary evaluators to evaluate technology students' written artifacts with acceptable levels of inter-rater reliability?
2) Do technology major students learn and demonstrate stronger skills in Introductory Composition (specifically research writing skills) when enrolled in an integrated interdisciplinary learning program?

RESEARCH METHODS

This research design was quasi-experimental, comprising two distinct groups of student artifacts. Writing samples were collected after the semester concluded from both integrated and non-integrated Introductory Composition courses. Integrated sections were those in which students were concurrently enrolled in the

Design Thinking in Technology course; non-integrated sections were those offered outside of the Integrated First-Year Experience program.

DATA COLLECTION

This study has been approved by the Institution Review Board (IRB) at Purdue University. Data were collected and analyzed with consent from students and instructors. The student writing samples from both integrated and non-integrated classrooms were collected from multiple sections of Introductory Composition offered during the Fall 2016 semester. The total was ninety-one samples from both integrated and non-integrated composition classrooms. Thirty-nine student artifacts (i.e., Research papers, Posters, Kickstarter documents) were collected from four of the six integrated English Composition sections. The data from two sections were not reviewed because of the nature of the assignments in those sections was not related to research or argumentation, as the others more clearly were. After collecting data, an analogous set of non-integrated English composition instructors were identified. Instructor similarity is relevant to this study since instructor expertise and experience often have an impact on student learning. Additionally, the Assistant Director of the first-year writing program at Purdue University was consulted for identifying non-integrated sections. Her involvement in the mentoring of first-time instructors meant that she knew instructors' teaching experience, backgrounds, and approaches. The Assistant Director recommended six non-integrated sections/instructors that were appropriately comparable to the six integrated sections/instructors for the current study. The excluded group for this research were specialized sections of English tailored toward learning communities, international students, and sections taught by first-time instructors. From the sections recommended, four instructors responded and consented to participate in the study. Therefore, all integrated and non-integrated sections included in this study were taught by instructors with comparable levels of experience (between 2 and 4 years teaching first-year composition at this institution).

Though instructor differences, as well as differences in students' academic preparedness, class standing, and experience were all controlled to the extent possible, it became apparent during the course of this research that not all students entered the courses with the same capabilities, backgrounds, and experiences. After data collection, it was also discovered that students in non-integrated English sections had entered Purdue University with higher overall ACT and/or SAT scores than students in integrated sections. Table 1. shows a comparison of composition students' SAT and/or ACT scores by group (integrated or non-integrated). Students in the integrated sections show lower SAT and ACT scores on all subscales. In particular and most critical to this study, students in the integrated group had significantly lower SAT Writing or ACT English scores before entering the university.

Students in non-integrated sections this semester had generally earned more credit hours at the university than students in integrated sections. About 80% of the students in the four non-integrated composition sections were categorized as freshmen. The integrated sections, by design, are intended for first-year students in their freshman year, and nearly 96% of students in all six integrated sections were counted as freshman by credit hours (see Table 2). Since the study was conducted in the fall semester, it is likely that most students in integrated sections were first-semester freshman or first-semester sophomores. It is likely that the 12 sophomores in non-integrated sections would have 3 semesters of academic experience, whereas the freshman only have 1 semester of experience. A chi-square test of goodness-of-fit was performed to determine

Table 1. English composition students' standardized test means (standard deviations).

Sections	SAT Verbal	SAT Math	SAT Writing	ACT Composite	ACT English	ACT Combined
Non-integrated	583.00 (64.70) N = 40	625.00 (90.70) N = 40	580.25** (79.95) N =40	27.47** (3.42) N = 38	26.89** (4.88) N = 38	25.86** (3.57) N = 29
Integrated	562.96 (69.58) N = 71	604.08 (69.58) N = 71	529.86** (60.08) N = 71	25.34** (2.82) N = 59	24.47** (3.59) N = 59	22.87** (25.86) N = 30)

Note: ** *p* < .01

Table 2. Distribution of composition students' class standings.

Sections	Freshmen	Sophomores	Juniors	Seniors
Non-integrated sections	45 (78%)	12 (21%)	0	1 (2%)
Integrated sections	93 (95%)	4 (4%)	1 (1%)	0

whether the distribution of class standings was similar across the two study groups. Differences between the integrated and non-integrated sections were significantly different, X^2 (3, *N* = 201) = 16.226, *p* =.001. Some differences in maturity and experience could have affected students' writing and composition skills as demonstrated in the artifacts that were collected.

These variations between student populations, some of which were statistically significant, make it difficult to parse out differences attributable to the effects of the two learning experiences from differences attributable to the groups' pre-existing differences. Though lower writing scores on standardized tests and potential differences in maturity will limit the conclusions and may mask discoveries related to the second research question about the Integrated First-Year Experience program, this study may serve as a jumping off point for future work. The preliminary results and conclusions drawn here about the first research question may also have broader implications.

Interdisciplinary Evaluators

Three evaluators from two different disciplinary backgrounds were involved in the project. They were: one Associate Professor in Technology, one advanced English PhD candidate with seven years' English teaching experience, and one first-year international English PhD student with only one year of teaching experience. One of the opportunities provided by ACJ is the inclusion of interdisciplinary judges, as Pollitt (2012) recognized. He argued, "there is no need to limit judgment to teachers. Any interested party could, in principle, be invited to make those judgments—to try the system and so gain a better understanding of its strengths" (Pollitt, 2012, p. 297). In the context of this research, all three evaluators were instructors, but from varied cultural and disciplinary backgrounds and with different levels of teaching experience at Purdue University.

Holistic Statement

Adaptive Comparative Judgment leverages a holistic statement to focus judges on key elements of learning to be assessed (Pollitt, 2012). Accordingly, the following holistic statement was used when making comparison between artifacts: "Which artifact demonstrates stronger writing and more rhetorical awareness?" This statement was developed based on the stated programmatic outcomes of Purdue University's Introductory Composition program. The statement is also based on the concept of rhetorical awareness and situated in a shared understanding of the rhetorical triangle (author, audience, and message) and key rhetorical appeals (ethos, logos, and pathos) ("Introductory Composition at Purdue University," n.d.). To clarify this brief holistic statement and to guide the ACJ comparison process, a one-page explanation was written by an experienced English Composition instructor who had been part of the pedagogical team behind the program's initial implementation. The statement reinforces the idea of holisticism as described by White (1984), and this research uses ACJ as a tool of assessment developed by Pollitt (2012) to assess multimodal student writing.

EVALUATION

The first evaluation began with a small pilot study in order to gauge the potential feasibility of using ACJ and to acclimate researchers to the ACJ interface and process. With minimal discussion about the student artifacts and no calibration efforts among the evaluators, an evaluation of 10 randomly selected student artifacts was conducted. With 10 samples, 12 rounds of judgment using ACJ were made, resulting in an evaluators' consistency coefficient (similar to inter-rater reliability) of 0.85 (see Pollitt, 2012; 2015 for an in-depth discussion of the evaluator consistency coefficient.) From the pilot study, estimated time commitment, potential technical issues in the interface, and other user-friendly techniques were all identified and planned for.

Following the pilot study, the full data set was compared without discussion, follow-up, or any other training among evaluators. The CompareAssess interface, which relies on the adaptive algorithm developed by Pollitt (2012) to facilitate the comparison, was utilized for 12 rounds of judgment, signifying that each artifact had been compared at least 12 times to another artifact. Following the initial 12 rounds, a check into the reliability of judgments was made and it was determined that an additional 4 rounds may improve the overall judgment and rank consistency. This follows the rationale behind ACJ, as multiple rounds of judgment are required to ensure fidelity of results; with multiple rounds, each artifact can be compared with several others by several judges, so that an accurate rank order can be produced (Pollitt, 2012; Rangel-Smith & Lynch, 2018).

RESULTS

In total, over the span of about one month, each evaluator completed approximately 250 comparisons (for a total of 16 rounds), leading to a total of 728 comparative judgments of 91 student papers. Data from the comparison of these artifacts, which were drawn from both integrated and non-integrated sections of English Composition and evaluated by three evaluators from differing backgrounds, became the foundation for answering the two research questions.

ACJ Reliability

The first research question addressed a pragmatic question about interdisciplinary education and evaluation by evaluators from different disciplines. The potential value of having different perspectives reviewing student work is that they may see different aspects of the composition as valuable, according to their fields' priorities. However, if these different values result in disagreement about which submissions are higher quality, the validity and reliability of the assessment methods will be low and the assessment questionable. The judge consistency coefficient (referred to as the "JCC" by Pollitt [2015]) is a measure of consistency in judgments between raters, that is, a form of inter-rater reliability. In this research, the JCC was used as an indicator of agreement between judges and, after 16 rounds of judgments, the JCC value was r = 0.71. This value was determined sufficient for this exploratory project; as Nunnally (1978) pointed out, a reliability of r = 0.70 or higher will suffice in the early or exploratory stages of a study. Recognizing the limits associated with the JCC reliability value obtained here, it was nevertheless determined that given the exploratory nature of this research, the collected data and results warranted further analysis and investigation as a potential catalyst for further inquiry and investigation.

Technology Majors' Learning in Integrated Courses

The second research question was contingent on the first and addressed differences in student learning between sections of Composition integrated with the Design Thinking in Technology course and sections of Composition not integrated with the design course. To answer this question, the data from the 16 rounds of ACJ judgments, including the parameter values obtained from the ACJ session (described previously) and the categorization variable (integrated vs. non-integrated) for each student paper were analyzed using a single sample t-test to investigate the difference in achievement between students in each of the treatment conditions. In this case, a higher parameter value means that a sample was consistently chosen as the "stronger" of the ACJ pairs, whereas a lower parameter value means that a sample was consistently not chosen. The results from the data analysis were significant: students in the integrated sections received significantly lower parameter values (M = -.42, SD = 1.12) than those in the non-integrated sections (M = .43, SD = 1.01); $t(97) = 4.00$, $p < .001$. The findings indicate that student submissions from the non-integrated sections were regarded by the ACJ judges as significantly better than those in the integrated sections.

Implications and Recommendations

The ACJ assessment method appears promising for the current culture of interdisciplinary writing and assessment. This study has demonstrated that three evaluators from different backgrounds could evaluate English composition assignments with an acceptable reliability. This is significant as interdisciplinary and integrated curriculum approaches are becoming more commonplace. Teaching from an interdisciplinary perspective requires assessment from an interdisciplinary perspective and this method was successful for this research. Measuring success and quality of composition work as it becomes dynamically entangled within other disciplines is increasingly complex. Future studies may replicate this work with larger and more diverse pools of evaluators or in other interdisciplinary contexts.

Evaluation of student work revealed significant differences between the integrated and non-integrated sections. However, it is important to note that while the findings indicated significantly lower-rated work from students in the integrated sections, these students also started out with significantly lower scores related to key English skills than their peers. The results of the research highlight the need for future studies which investigate students' English writing skills and establish equivalency prior to the beginning of the study. Alternatively, in order to establish comparable populations, researchers could create a baseline for both populations and then measure growth from there. In this research, the section selections were based on comparable instructor skills and experience, not based on comparable students as that student data was not readily available at the onset of the study.

Another limitation of the study was that in the integrated sections, assigned research papers did not consistently align with students' integrated projects and topics as expected. This meant that using these artifacts to test whether more learning resulted from the programmatic integration may not be as valid as anticipated. It is possible that in some cases, less time and emphasis may have been placed on the research assignment in the integrated sections, in accommodation of the extra effort spent on the larger integrated project. This possibility may potentially be masking any more significant learning gains which may or may not have existed.

As a result of these research efforts, two implications for other technology programs can be suggested. First, while literature suggests that students may learn more effectively in integrated programs where learning across multiple disciplines is explicitly connected, this may not be the case. The results from this study also suggest that making explicit integrations between the design thinking course and the composition and communication courses may not necessarily improve student learning. Therefore, the suggestions from this work include future experimentation with integration among courses to measure impacts on student learning. Second, as a result of success in using ACJ, this research suggests technology programs enlist judges from the core content areas of technology along with judges from composition and communication backgrounds to integrate evaluation efforts. Including interdisciplinary perspectives in program assessment efforts is one important way of accommodating technology programs to hirers' increasingly complex demands for innovative, critical technology graduates.

CONCLUSION

This research experimented with using evaluators from different disciplines to assess student learning in an integrated interdisciplinary environment compared to a traditional approach using ACJ. In the future, the researchers should aim at conducting a study of English Composition learning where the treatment group and comparison group have been more intentionally selected, factoring in test scores and student demographics, so that both populations are more similar. Determining that students in both integrated and non-integrated sections of English Composition have comparable skills at the onset of the semester will yield a more rigorous study.

Based on the positive experience with interdisciplinary evaluators, the study concludes that further experimentation with an interdisciplinary evaluation team in analyzing student submissions to help meet the current demands of English education is warranted. Based on the findings from the comparison of student submissions, the research recommends additional research into the integrated relationships between academic disciplines which may or may not be beneficial for technology major students, as suggested by the National Academy of Engineering and the National Research Council.

Sweta Baniya *is a PhD Candidate in rhetoric and composition at Purdue University, West Lafayette, Indiana.*

Nathan Mentzer *is an Associate Professor of Engineering and Technology Teacher Education at Purdue University, West Lafayette, Indiana*

Scott R. Bartholomew *is an Assistant Professor of Engineering/Technology Teacher Education at Purdue University, West Lafayette, Indiana.*

Amelia Chesley *is an Assistant Professor in the Department of English, Foreign Languages, and Cultural Studies at Northwestern State University of Louisiana, Natchitoches.*

Cameron Moon *is a graduate student teaching assistant for the Technology, Leadership and Innovation department at Purdue University, West Lafayette, Indiana.*

Derek R. Sherman *is a PhD Candidate in rhetoric and composition at Purdue University, West Lafayette, Indiana.*

REFERENCES

Bannerot, R., Kastor, R., & Ruchhoeft, P. (2010). Multidisciplinary capstone design at the University of Houston. Advances in Engineering Education 2(1), 1–33.

Bartholomew, S. R., Strimel, G. J., & Yoshikawa, E. (2018). Using adaptive comparative judgment for student formative feedback and learning during a middle school design project. *International Journal of Technology and Design Education.* https://doi.org/10.1007/s10798-018-9442-7

Bartholomew, S. R., & Yoshikawa, E. (2018). A systematic review of research around Adaptive Comparative Judgment (ACJ) in K-16 education. 2018 CTETE Monograph Series, https://doi.org/10.21061/ctete-rms.v1.c.1

CCCCs Position Paper on Writing Assessment: "A Statement on Education Issue approved by the Executive Committee," 2014. Retrieved From https://cccc.ncte.org/cccc/resources/positions/writingassessment

Gallagher, C. W. (2014). Immodest witnesses: Reliability and writing assessment. *Composition Studies*, 42(2), 73-95.

Huot, B. A. (2002). *Rearticulating writing assessment for teaching and learning.* Logan, Utah: Utah State Press.

Honey, M, Pearson, G, Schweingruber, H. (2014) *STEM Integration in K-12 Education: Status, Prospects, and an Agenda for Research.* Washington, DC: The National Academies Press

Introductory Composition at Purdue, n.d. Retrieved from https://icap.rhetorike.org/outcomes/

Kellam, N., Walther, J., Costantino, T., Dodd L., & Cramond, B. (2013, Winter). Integrating the engineering curriculum through the synthesis and design studio. *Advances in Engineering Education*, 1–33.

Lynne, P. (2004). *Coming to terms: A Theory of writing assessment.* Logan, Utah: State University Press.

Mansilla, V. B., Duraisingh, E. D., Wolfe, C. R., Haynes, C. (2009). Targeted assessment rubric: An empirically grounded rubric for interdisciplinary writing. *The Journal of Higher Education*, 80(3), 334-353.

Moss, P. A. (1994). Validity in high stakes writing assessment: Problems and possibilities. *Assessing Writing*, 1(1), 109-128.

National Academy of Engineering, (n.d.). Retrieved From: http://www.engineeringchallenges.org/

Nunnally, J. C. (1978). *Psychometric theory* (2nd ed.). New York: McGraw-Hill.

Pollit, A (2012). The method of adaptive comparative judgment. *Assessment in Education: Principles, Policy & Practice*, 19(3), 281-300.

Pollitt, Alastair. (2015). On 'Reliability' bias in ACJ. *Cambridge Exam Research.* 10.13140/RG.2.1.4207.3047.

Rangel-Smith, C. & Lynch, D. (2018). Addressing the issue of bias in the measurement of reliability in the method of Adaptive Comparative Judgment. Paper presented at the 36th Pupils' Attitudes towards Technology Conference, Athlone, Ireland, pp. 378-387

Steedle, J. T., & Ferrara, S. (2016). Evaluating comparative judgment as an approach to essay scoring. *Applied Measurement in Education*, 29(3), 211–223. https://doi.org/10.1080/08957347.2016.1171769

Van Daal, T., Lesterhuis, M., Coertjens, L., Donche, V., & De Maeyer, S. (2019). Validity of comparative judgement to assess academic writing: examining implications of its holistic character and building on a shared consensus. *Assessment in Education: Principles, Policy & Practice*, 1–16. https://doi.org/10.1080/0969594X.2016.1253542

White, E. M. (1984). Holisticism. CCC, 34(4), 400-409.

Wang, H., Moore, T., Roehrig, G. H., & Park, Mi Sun. (2011). STEM integration: Teacher perceptions and practice. *Journal of Pre-College Engineering Education Research*, 1(2). Retrieved from http://dx.doi.org/10.5703/1288284314636

Yancey, K. B. Writing Assessment in the Early Twenty-first Century: A Primer. In Ritter, K & Matsuda, P.K. (2012) *Exploring Composition Studies: Sites, Issues, and Perspectives* (167-87). Utah: Utah State.